BITS F464 Machine Learning

Assignment 2

Date : 9th April, 2020

Team Members

Durba Satpathi	2019A7PS0972H
R Adarsh	2019A7PS0230H
Navdeep Singh	2017B5A71675H

Problem 2C Comprehensive Comparison

3.1. A comparative analysis of the models and their accuracies

Dataset given:
The given dataset consists of 18185 instances and each input instance has 10 attributes. Each instance belongs to a class Jasmine or Gonen.
Cross Validation:
The dataset is split into k(=7) groups (7-fold cross validation). For each unique group, consider it as the test set and remaining groups as a train set. Fit the model on the train set and evaluate on the test set and obtain the evaluation metrics. Finally we summarize the overall evaluation score using the sample of model evaluation metrics.

Logistic Regression Vs SVM:
SVM can handle non-linear solutions too whereas Logistic regression can handle linear data only. Linear SVM handles outliers better. LR performs well with linear data.

SVM Vs Naive Bayes:
Both perform better with less train data and large features, however if there is mutual dependency among features SVM outperforms Naive Bayes.

SVM Vs ANN:
SVM leads to convex optimization whereas ANN may run into local optima. ANN requires large training data to perform well whereas SVM can give good accuracy with less training data.

Logistic Regression Vs ANN:
ANN supports non-linear solutions whereas LR does not. Due to the convex nature of loss function LR does not run into local minima.

Logistic Regression Vs Naive Bayes:
Naive Bayes is a generative model whereas LR is a discriminative model. LR performs better than NB when there is dependence between features.

FLD and **Linear perceptron** work good when the data is linearly separable. If the data is linearly separable, Perceptron always converges to minimize the loss and achieve 100% accuracy.

Training Accuracy
Naive Bayes
Fold 0 Accuracy- 0.9830638953040801 F1 Score- 0.9818031430934656
Fold 1 Accuracy- 0.9872979214780601 F1 Score- 0.9854175872735308

Fold 2 Accuracy- 0.9892224788298691 F1 Score- 0.9879621668099742
Fold 3 Accuracy- 0.9861431870669746 F1 Score- 0.9840989399293285
Fold 4 Accuracy- 0.9815242494226328 F1 Score- 0.9791485664639444
Fold 5 Accuracy- 0.9819091608929946 F1 Score- 0.9798887462558836
Fold 6 Accuracy- 0.9826723142087024 F1 Score- 0.9807445442875481
0.9845476010290449

Logistic Regression
Fold 0 Accuracy- 0.9892224788298691 F1 Score- 0.9884963023829089
Fold 1 Accuracy- 0.9896073903002309 F1 Score- 0.9881526985519964
Fold 2 Accuracy- 0.9919168591224018 F1 Score- 0.9910371318822023
Fold 3 Accuracy- 0.9911470361816782 F1 Score- 0.9898989898989898
Fold 4 Accuracy- 0.9903772132409546 F1 Score- 0.989172802078822
Fold 5 Accuracy- 0.9876828329484219 F1 Score- 0.9863597612958226
Fold 6 Accuracy- 0.9899884482094725 F1 Score- 0.9889830508474577
0.9899917512618613
Artificial Neural Networks
Fold 0 Accuracy- 0.9876828329484219 F1 Score- 0.9868529170090385
Fold 1 Accuracy- 0.9896073903002309 F1 Score- 0.9881109643328929
Fold 2 Accuracy- 0.9896073903002309 F1 Score- 0.9885447602885022
Fold 3 Accuracy- 0.99153194765204 F1 Score- 0.9903593339176161
Fold 4 Accuracy- 0.9892224788298691 F1 Score- 0.9878682842287695
Fold 5 Accuracy- 0.9842186297151655 F1 Score- 0.9826491747778248
Fold 6 Accuracy- 0.9896033885252215 F1 Score- 0.9885835095137421
0.9887820083244544

FLD
Fold 0 Accuracy- 0.9830638953040801 F1 Score- 0.9816971713810316
Fold 1 Accuracy- 0.985373364126251 F1 Score- 0.9831261101243339
Fold 2 Accuracy- 0.9880677444187836 F1 Score- 0.9866206301251619
Fold 3 Accuracy- 0.985373364126251 F1 Score- 0.9831261101243339
Fold 4 Accuracy- 0.9830638953040801 F1 Score- 0.9807186678352323
Fold 5 Accuracy- 0.9807544264819091 F1 Score- 0.9784853700516352
Fold 6 Accuracy- 0.9826723142087024 F1 Score- 0.9806784027479606
0.9840527148528654

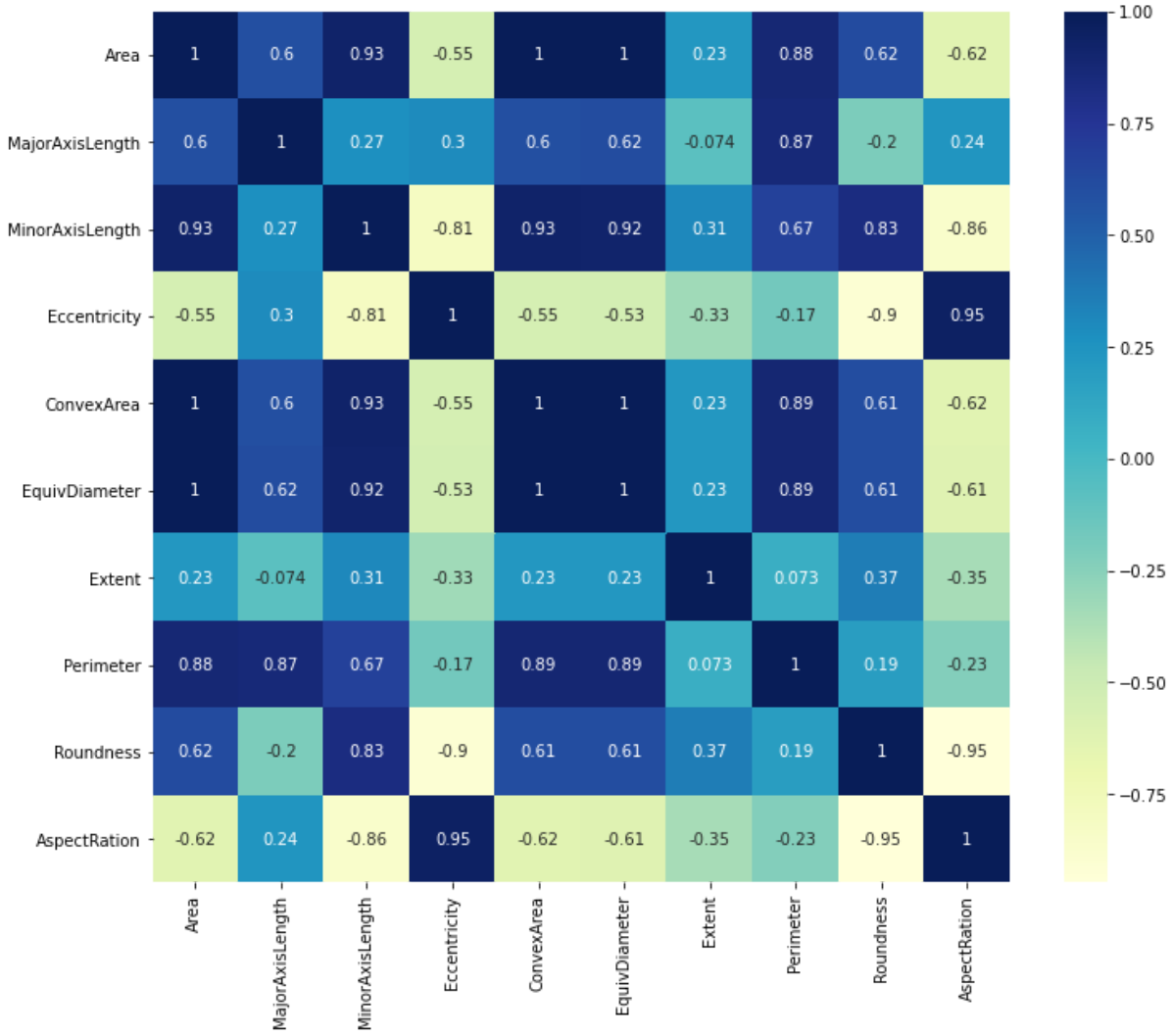
SVM
Fold 0 Accuracy- 0.9861431870669746 F1 Score- 0.9851607584501237
Fold 1 Accuracy- 0.9884526558891455 F1 Score- 0.9867957746478873
Fold 2 Accuracy- 0.9919168591224018 F1 Score- 0.9910371318822023
Fold 3 Accuracy- 0.9907621247113164 F1 Score- 0.9894366197183098
Fold 4 Accuracy- 0.9888375673595073 F1 Score- 0.9874295622019938
Fold 5 Accuracy- 0.9861431870669746 F1 Score- 0.9846547314578006
Fold 6 Accuracy- 0.9861378513669619 F1 Score- 0.9846938775510206
0.9883419189404689
Perceptron
Fold 0 Accuracy- 0.9846035411855273 F1 Score- 0.9835526315789472
Fold 1 Accuracy- 0.9738260200153964 F1 Score- 0.9693417493237151
Fold 2 Accuracy- 0.9911470361816782 F1 Score- 0.990183525394793
Fold 3 Accuracy- 0.9661277906081601 F1 Score- 0.9626485568760611
Fold 4 Accuracy- 0.985373364126251 F1 Score- 0.9833916083916084
Fold 5 Accuracy- 0.9849884526558892 F1 Score- 0.9834254143646409
Fold 6 Accuracy- 0.9715055833654216 F1 Score- 0.9677137870855148
0.979653112591189

Model	Training Accuracy	F Score	Recall	Precision	Testing accuracy
Naïve Bayes	0.9846	0.985	0.993	0.978	0.9845
Logistic Regression	0.9892	0.991	0.994	0.988	0.9897
ANN	0.9876	0.989	0.991	0.987	0.9891
SVM	0.9886	0.989	0.994	0.985	0.9884
LDA	0.9841	0.985	0.990	0.975	0.9842
Linear Perceptron	0.9793	0.984	0.976	0.985	0.9797

3.2. The model that performed best and one the that performed worst

Based on the metrics of the different models we see that all the models had accuracies above 97%. LR,ANN tend to perform slightly better than Naive Bayes,LDA and Perceptron.

- LR is the best performer. This might be because LR performs better when there is dependence between features and its convex loss function ensures that we don't run into local minima.
- From the correlation matrix, we see that many of the features are highly correlated and naive bayes assumes that the features are independent ,this might be one of the reasons for comparatively less accuracy of naive bayes.


Area 1 0.6 0.93 -0.55 1 1 0.23 0.88 0.62 -0.62
MajorAxisLength 0.6 1 0.27 0.3 0.6 0.62 -0.074 0.87 -0.2 0.24
MinorAxisLength 0.93 0.27 1 -0.81 0.93 0.92 0.31 0.67 0.83 -0.86
Eccentricity -0.55 0.3 -0.81 1 -0.55 -0.53 -0.33 -0.17 -0.9 0.95
ConvexArea 1 0.6 0.93 -0.55 1 1 0.23 0.89 0.61 -0.62
EquivDiameter 1 0.62 0.92 -0.53 1 1 0.23 0.89 0.61 -0.61
Extent 0.23 -0.074 0.31 -0.33 0.23 0.23 1 0.073 0.37 -0.35
Perimeter 0.88 0.87 0.67 -0.17 0.89 0.89 0.073 1 0.19 -0.23
Roundness 0.62 -0.2 0.83 -0.9 0.61 0.61 0.37 0.19 1 -0.95
AspectRation -0.62 0.24 -0.86 0.95 -0.62 -0.61 -0.35 -0.23 -0.95 1
Area
MajorAxisLength
MinorAxisLength
Eccentricity
ConvexArea
EquivDiameter
Extent
Perimeter
Roundness
AspectRation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

3.3. The image containing box-plots for each modele